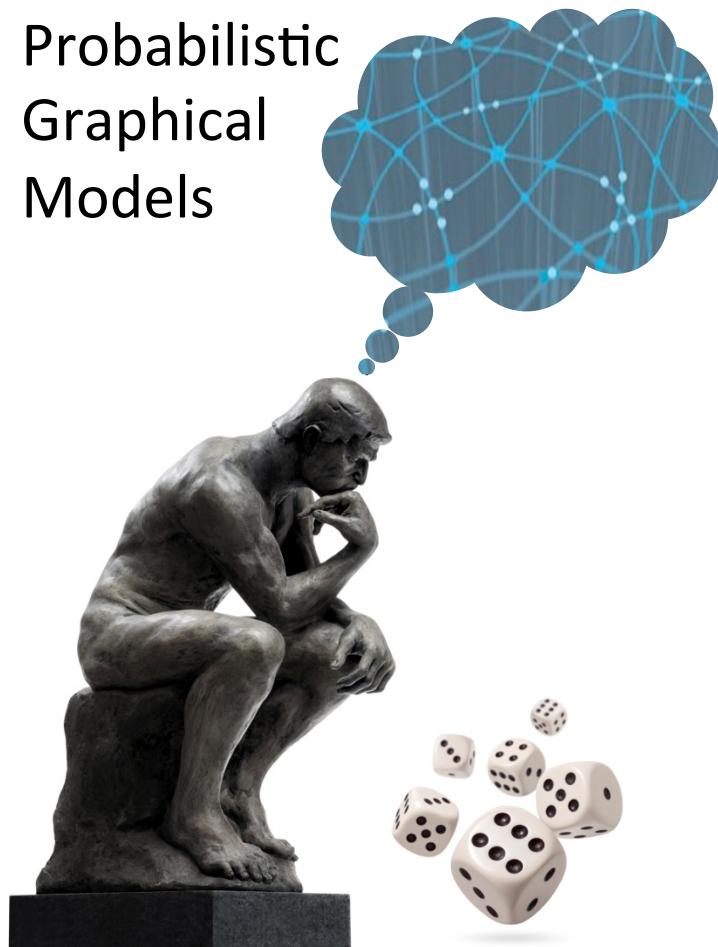


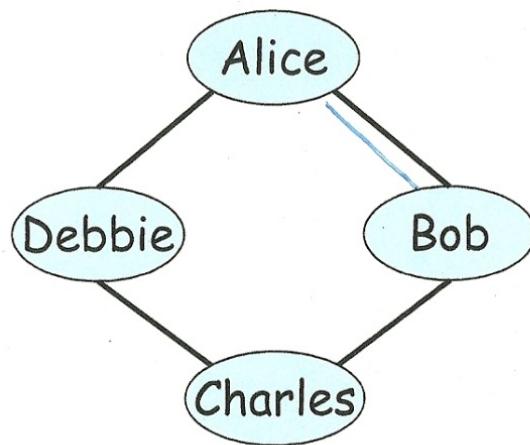
Probabilistic
Graphical
Models

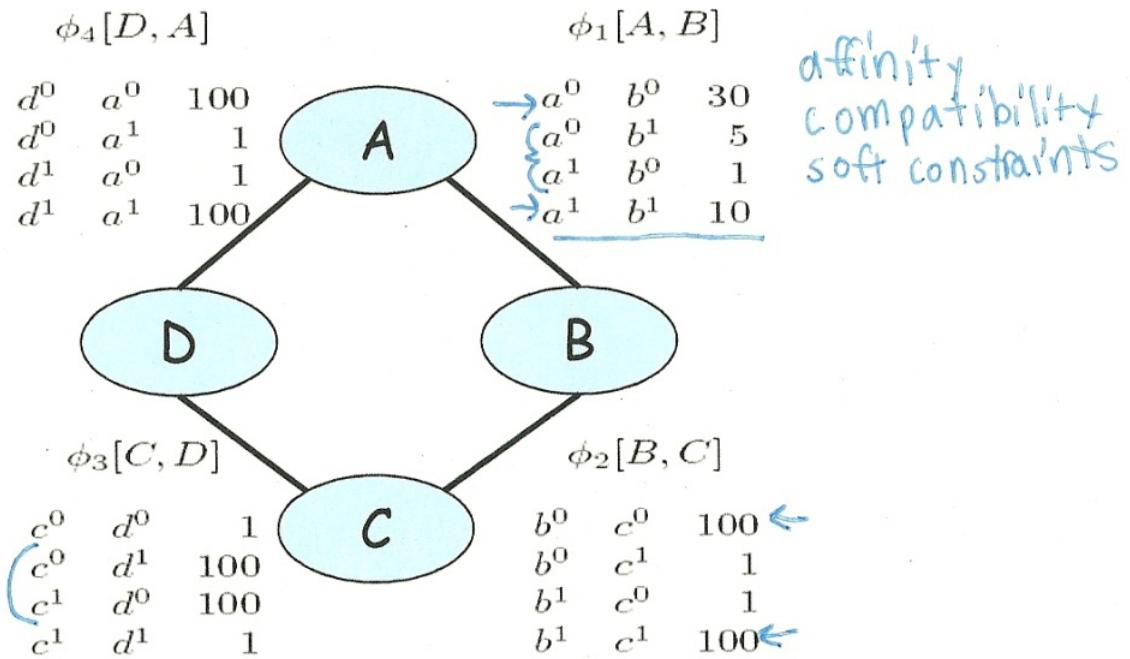


Representation

Markov Networks

Pairwise
Markov
Networks





$$\tilde{P}(A, B, C, D) = \phi_1(A, B) \times \phi_2(B, C) \times \phi_3(C, D) \times \phi_4(A, D)$$

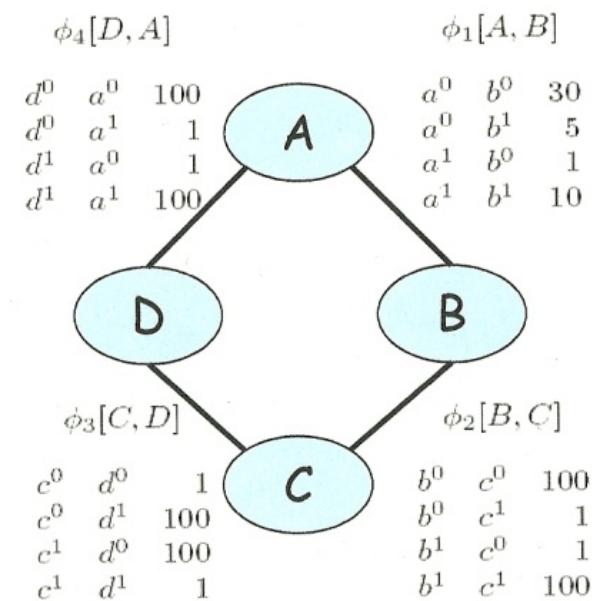
unnormalized measure

$$P(A, B, C, D) = \frac{1}{Z} \tilde{P}(A, B, C, D)$$

partition function

Assignment				Unnormalized
a^0	b^0	c^0	d^0	300000
a^0	b^0	c^0	d^1	300000
a^0	b^0	c^1	d^0	300000
a^0	b^0	c^1	d^1	30
a^0	b^1	c^0	d^0	500
a^0	b^1	c^0	d^1	500
a^0	b^1	c^1	d^0	5000000
a^0	b^1	c^1	d^1	500
a^1	b^0	c^0	d^0	100
a^1	b^0	c^0	d^1	1000000
a^1	b^0	c^1	d^0	100
a^1	b^0	c^1	d^1	100
a^1	b^1	c^0	d^0	10
a^1	b^1	c^0	d^1	100000
a^1	b^1	c^1	d^0	100000
a^1	b^1	c^1	d^1	100000

Z

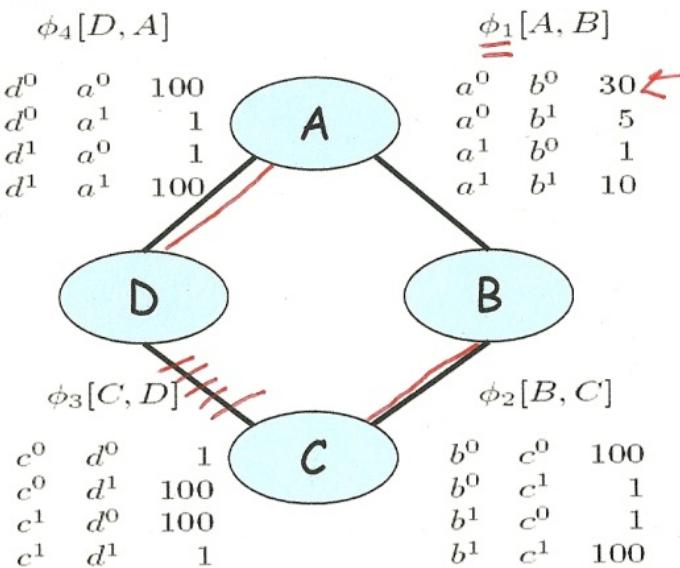


Daphne Koller

$p(A, B)$

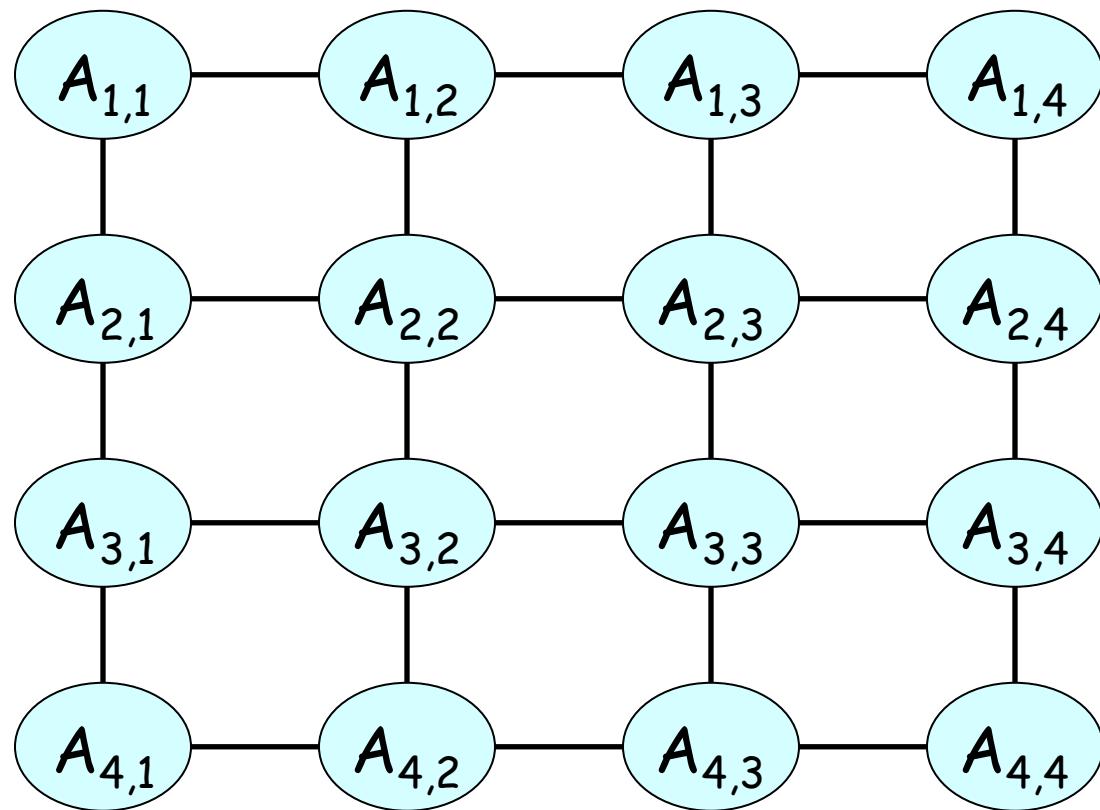
A	B	Prob.
a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

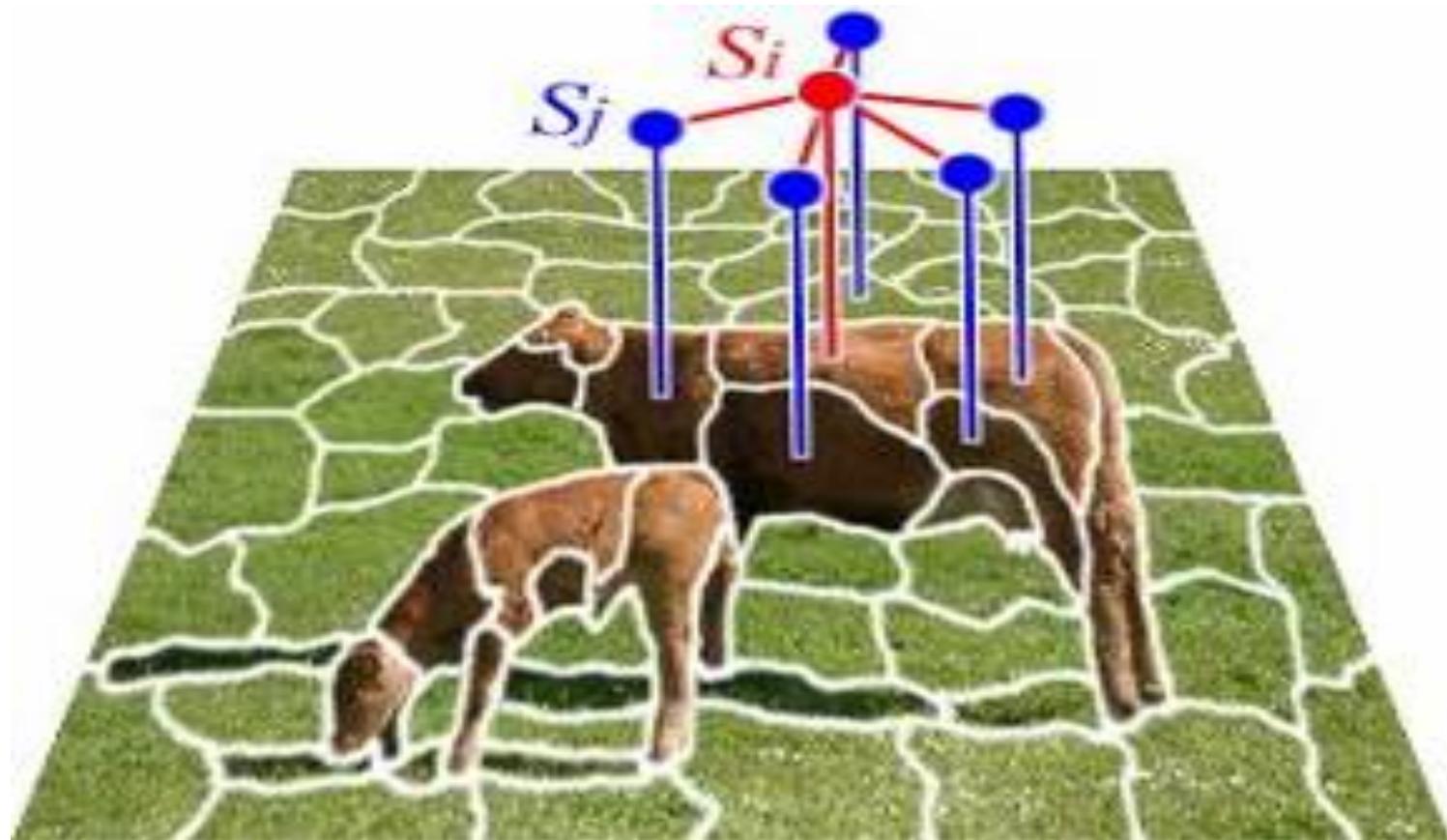
$$\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$$



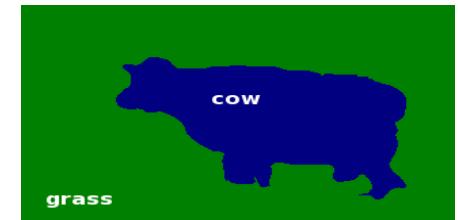
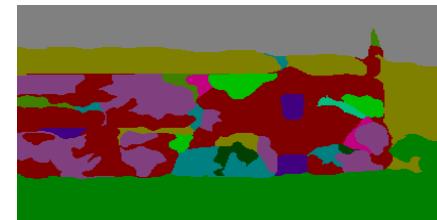
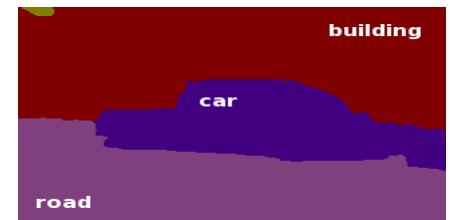
Pairwise Markov Networks

- A pairwise Markov network is an undirected graph whose nodes are X_1, \dots, X_n and each edge $\underline{X_i - X_j}$ is associated with a factor (potential) $\phi_{ij}(X_i, \overset{\text{random variables}}{X_j})$





Daphne Koller



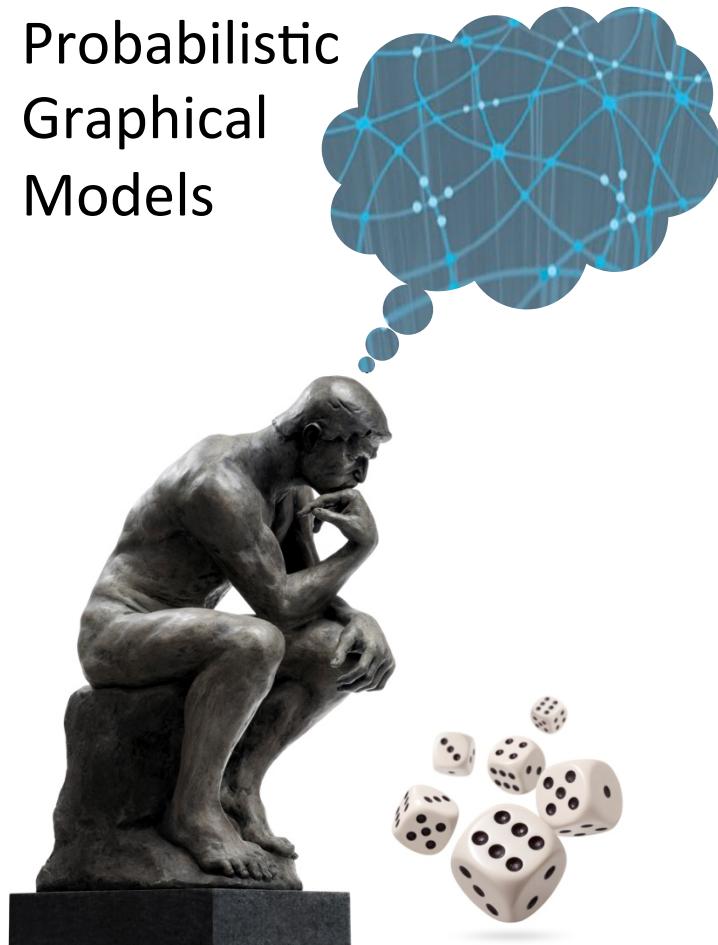
(a)

(b)

(c)

(d)

Probabilistic
Graphical
Models

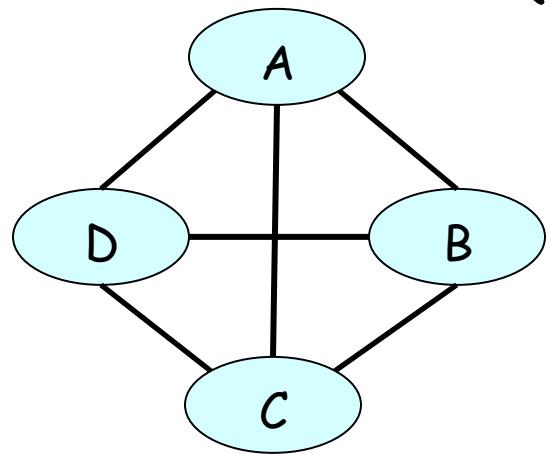


Representation

Markov Networks

General Gibbs
Distribution

$P(A, B, C, D)$



Is this fully expressive?

Gibbs Distribution

- Parameters:

General factors $\phi_i(D_i)$

$$\Phi = \{\phi_i(D_i)\}$$

a ¹	b ¹	c ¹	0.25
a ¹	b ¹	c ²	0.35
a ¹	b ²	c ¹	0.08
a ¹	b ²	c ²	0.16
a ²	b ¹	c ¹	0.05
a ²	b ¹	c ²	0.07
a ²	b ²	c ¹	0
a ²	b ²	c ²	0
a ³	b ¹	c ¹	0.15
a ³	b ¹	c ²	0.21
a ³	b ²	c ¹	0.09
a ³	b ²	c ²	0.18

Gibbs Distribution

Set of factors

$$\underline{\Phi} = \{\phi_1(D_1), \dots, \phi_k(D_k)\}$$

unnormalized measure k

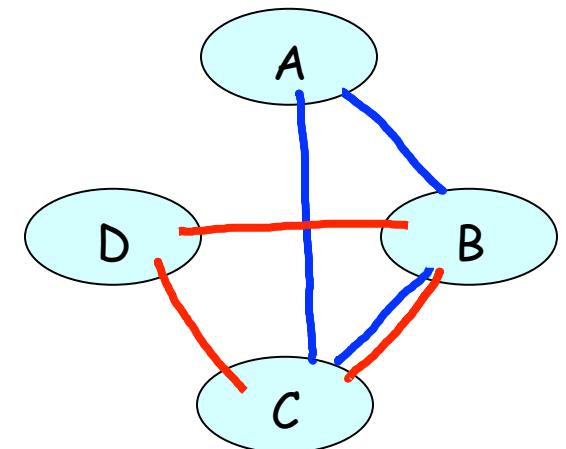
$$\tilde{P}_\Phi(X_1, \dots, X_n) = \prod \phi_i(\underline{D_i}) \quad \text{factor product}$$

$$Z_\Phi = \sum_{\underline{X_1, \dots, X_n}} \tilde{P}_\Phi(\underline{X_1}, \dots, X_n)$$

$$\overline{P}_\Phi(X_1, \dots, X_n) = \frac{1}{\underline{Z_\Phi}} \tilde{P}_\Phi(X_1, \dots, X_n)$$

Induced Markov Network

$\phi_1(\underline{A}, \underline{B}, C)$, $\phi_2(B, \underline{C}, D)$



$\Phi = \{\phi_1(D_1), \dots, \phi_k(D_k)\}$

Induced Markov network H_Φ has an edge $X_i - X_j$ whenever
there exists $\phi_m \in \Phi$ s.t. $x_i, x_j \in D_m$

Factorization

P factorizes over H if

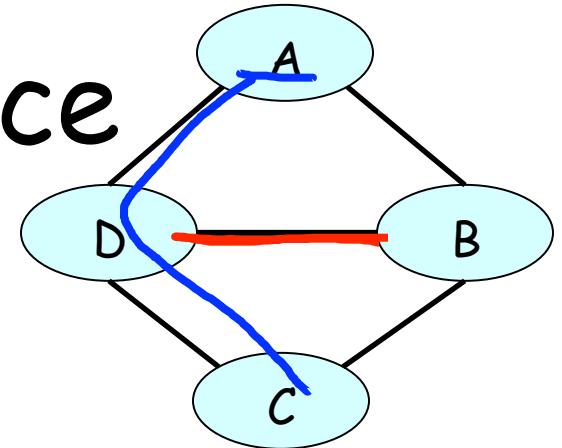
there exist $\underline{\Phi} = \{\phi_1(D_1), \dots, \phi_k(D_k)\}$

such that

$\underline{P} = P_{\underline{\Phi}}$ normalized product of
 \underline{H} is the induced graph for $\underline{\Phi}$ factors in $\underline{\Phi}$

Flow of Influence

$\phi_1(A, B, D)$, $\phi_2(B, C, D)$

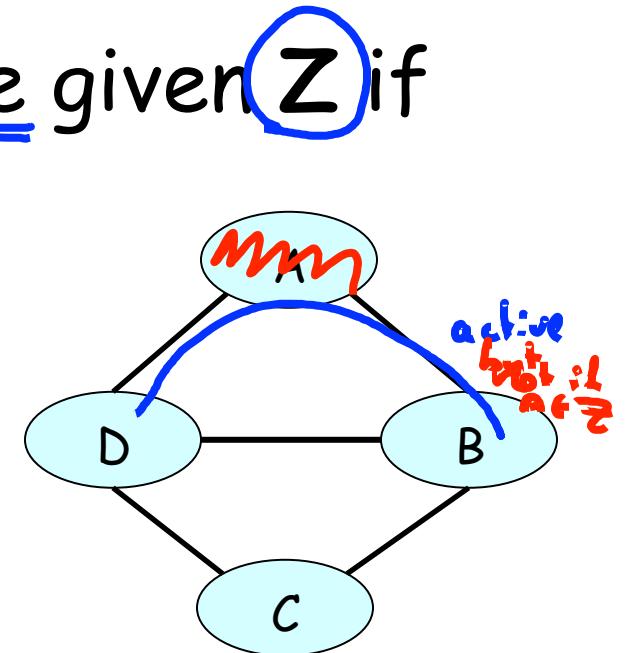


$\phi_1(A, B)$, $\phi_2(B, C)$, $\phi_3(C, D)$, $\phi_4(A, D)$, $\phi_5(B, D)$

- Influence can flow along any trail, regardless of the form of the factors

Active Trails

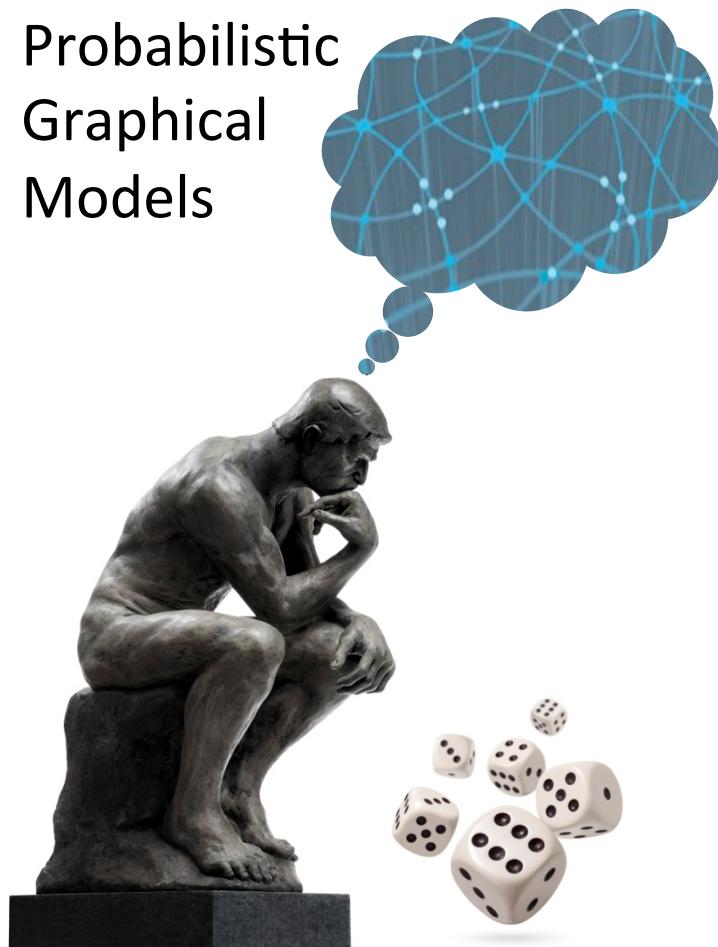
- A trail $X_1 - \dots - X_n$ is active given Z if no X_i is in Z



Summary

- Gibbs distribution represents distribution as a product of factors
- Induced Markov network connects every pair of nodes that are in the same factor
- Markov network structure doesn't fully specify the factorization of P
- But active trails depend only on graph structure

Probabilistic
Graphical
Models



Representation

Markov Networks

Conditional
Random
Fields

Motivation

- Observed variables X
- Target variables Y

$$X \rightarrow Y$$

$$P(X, Y)$$

joint

$$P(Y|X)$$

conditional

X

Y

image
segmentation

pixel values

pixel labels

text processing

words in a
sentence

parts of speech

CRF Representation

$$\phi_1(D_1), \dots, \phi_k(D_k)$$

$$\tilde{P}(\bar{x}, \bar{y}) = \prod_{i=1}^k \phi_i(D_i) \quad \text{unnormalized measure}$$

$$z(\bar{x}) = \sum_{\bar{y}} \tilde{P}(\bar{x}, \bar{y}) \quad \text{different } \tilde{z}(\bar{x}) \text{ for every assignment to the obs. variables } \bar{x}$$

$$P(\bar{y} | \bar{x}) = \frac{1}{z(\bar{x})} \tilde{P}(\bar{x}, \bar{y}) \quad \left(\sum_{\bar{y}} P(\bar{y} | \bar{x}) = 1 \text{ for all } \bar{x} \right)$$

CRFs and Logistic Model

$$\phi_i(X_i, Y) = \exp\{w_i \mathbf{1}\{X_i = 1, Y = 1\}\}$$

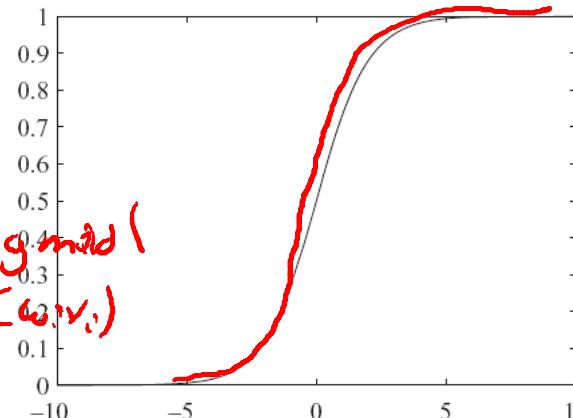
$$\phi_i(x_i, Y=1) = \exp\{\omega_i x_i\}$$

$$\phi_i(x_i, Y=0) = 1$$

$$\tilde{P}(Y=1 | X_1, \dots, X_n) = \exp\left(\sum_i \omega_i x_i\right)$$

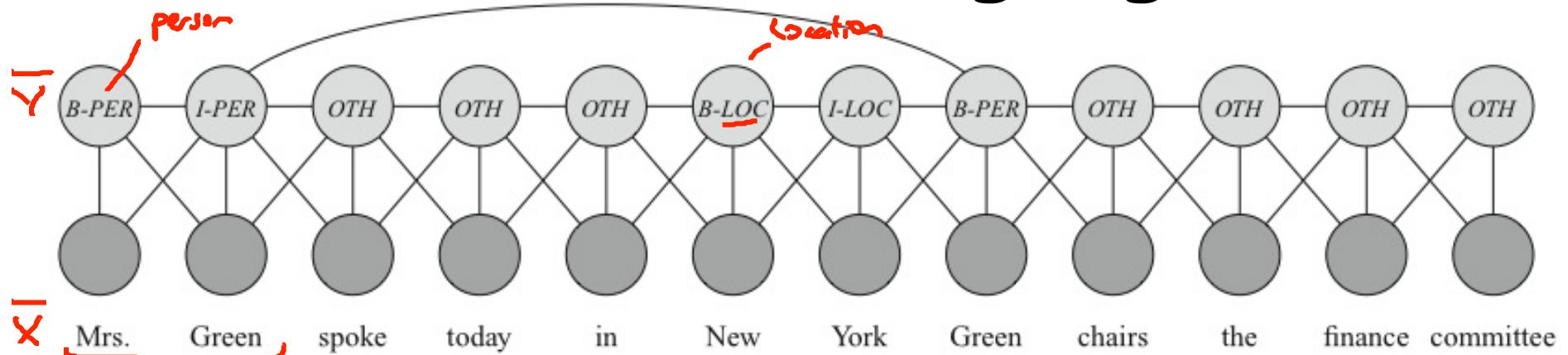
$$\tilde{P}(Y=0 | X_1, \dots, X_n) = 1$$

$$P(Y=1 | X_1, \dots, X_n) = \frac{\exp\left(\sum_i \omega_i x_i\right)}{1 + \exp\left(\sum_i \omega_i x_i\right)} = \text{sigmoid}\left(\sum_i \omega_i x_i\right)$$



Daphne Koller

CRFs for Language

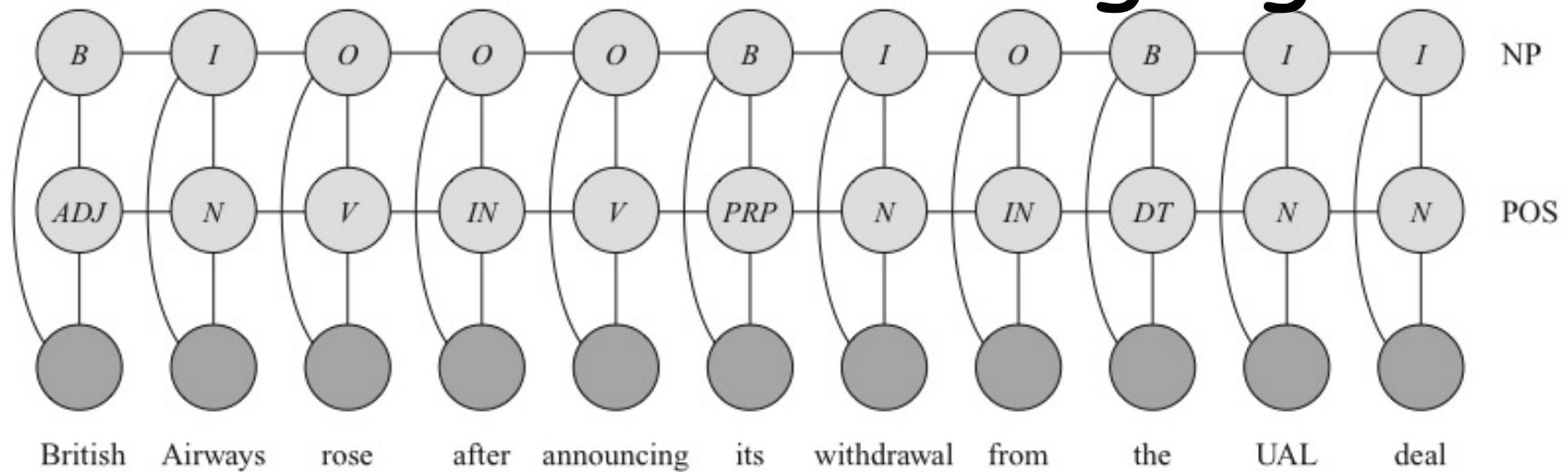


Features: word capitalized, word in atlas or name list, previous word is "Mrs", next word is "Times", ...

$$\tilde{P}(\vec{x}, \vec{y})$$

$$\Rightarrow P(\vec{Y}|\vec{X})$$

More CRFs for Language



KEY

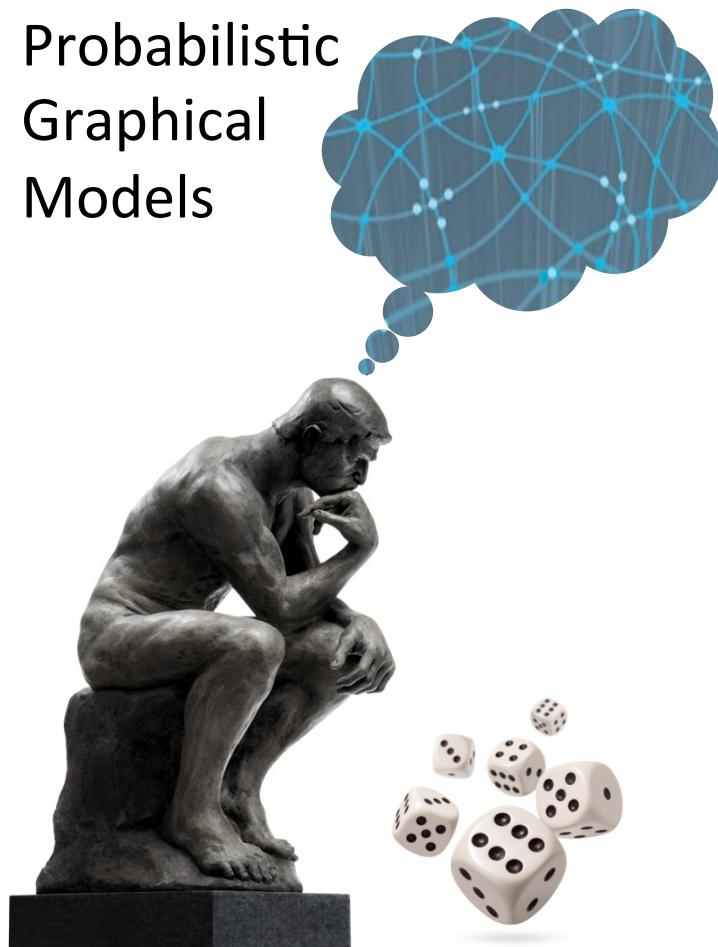
<i>B</i>	Begin noun phrase	<i>V</i>	Verb
<i>I</i>	Within noun phrase	<i>IN</i>	Preposition
<i>O</i>	Not a noun phrase	<i>PRP</i>	Possessive pronoun
<i>N</i>	Noun	<i>DT</i>	Determiner (e.g., a, an, the)
<i>ADJ</i>	Adjective		

Daphne Koller

Summary

- A CRF is parameterized the same as a Gibbs distribution, but normalized differently
- Don't need to model distribution over variables we don't care about
- Allows models with highly expressive features, without worrying about wrong independencies

Probabilistic
Graphical
Models



Representation

Independencies

Markov
Networks

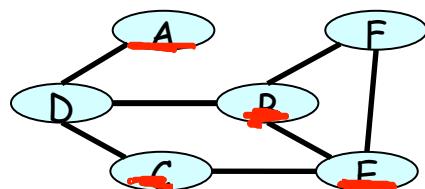
Separation in MNs

Definition:

X and Y are separated in H given Z

if there is no active trail in H

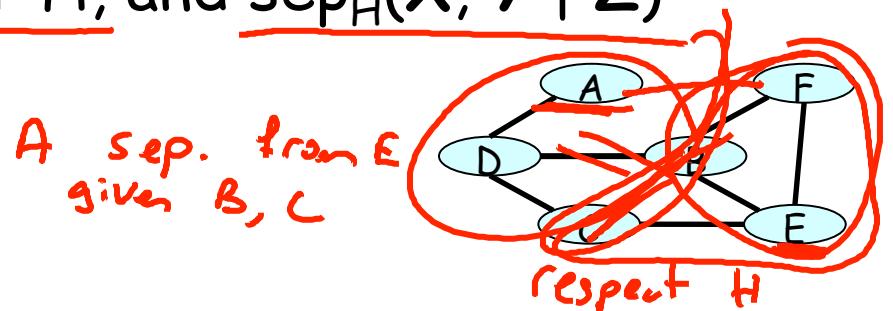
between X and Y given Z *no node along trail is in z*



A,E separated given B,D
given D
given B,C

Factorization \Rightarrow Independence: MNS

Theorem: If P factorizes over H, and $\text{sep}_H(X, Y | Z)$
then P satisfies $(X \perp Y | Z)$



$$\pi \phi_{\text{on } A} \cdot \pi \phi_{\text{on } E} =$$

cannot involve E cannot involve A

Factorization \Rightarrow Independence: MNs

$$\underline{I(H) = \{(X \perp Y \mid Z) : \boxed{\text{sep}_H(X, Y \mid Z)}\}}$$

If P satisfies I(H), we say that H is an I-map
(independency map) of P

Theorem: If P factorizes over H, then H is an I-map of P

Independence \Rightarrow Factorization

- Theorem (Hammersley Clifford):

For a positive distribution P , if H is an I-map for P , then P factorizes over H

$$P(z) \propto \alpha_z$$

Summary

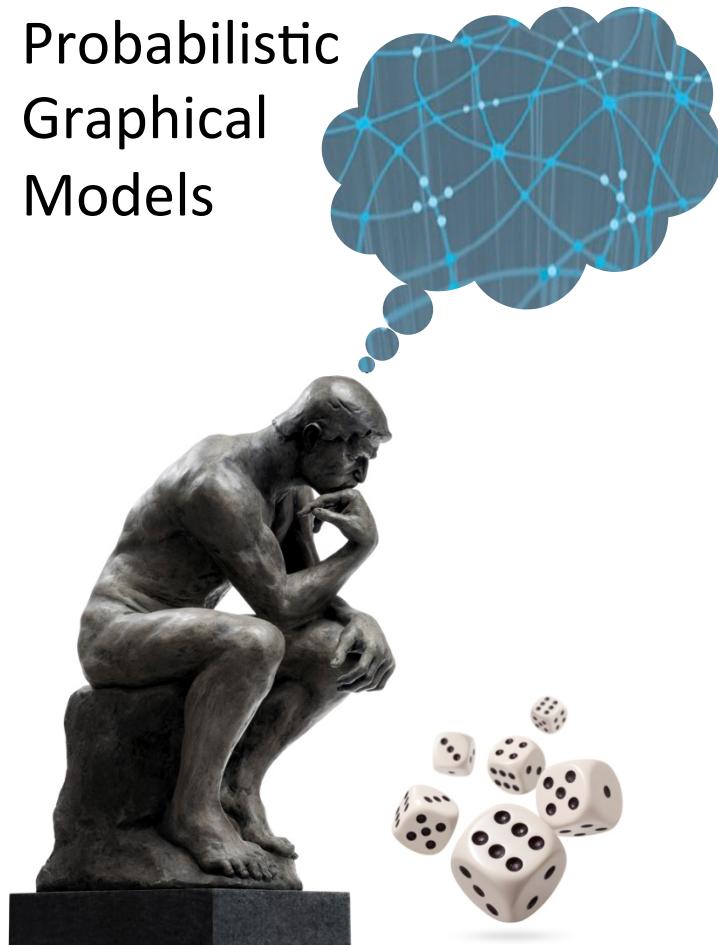
Two equivalent* views of graph structure:

- Factorization: H allows P to be represented
- I-map: Independencies encoded by H hold in P

If P factorizes over a graph H , we can read from the graph independencies that must hold in P (an independency map)

* for positive distributions

Probabilistic
Graphical
Models



Representation

Independencies

I-maps and
Perfect Maps

Capturing Independencies in P

$$\underline{I(P)} = \{\underbrace{(X \perp Y \mid Z)}_{\text{d-separation}} : \underline{P} \models \underbrace{(X \perp Y \mid Z)}_{\text{independencies that hold in } P}\}$$

- P factorizes over G \Rightarrow G is an I-map for P:

$$\text{d-separation} \quad \underline{I(G)} \subseteq I(P)$$

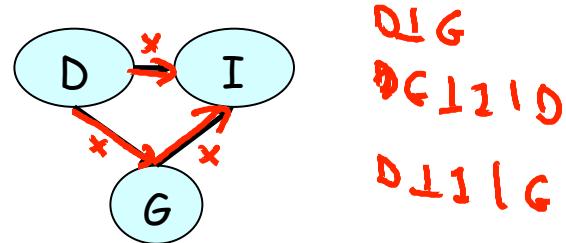
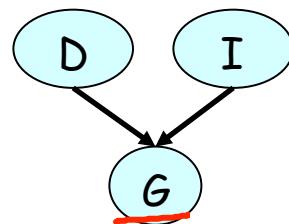
- But not always vice versa: there can be independencies in I(P) that are not in I(G)

Want a Sparse Graph

- If the graph encodes more independencies
 - it is sparser (has fewer parameters)
 - and more informative
- Want a graph that captures as much of the structure in P as possible

Minimal I-map

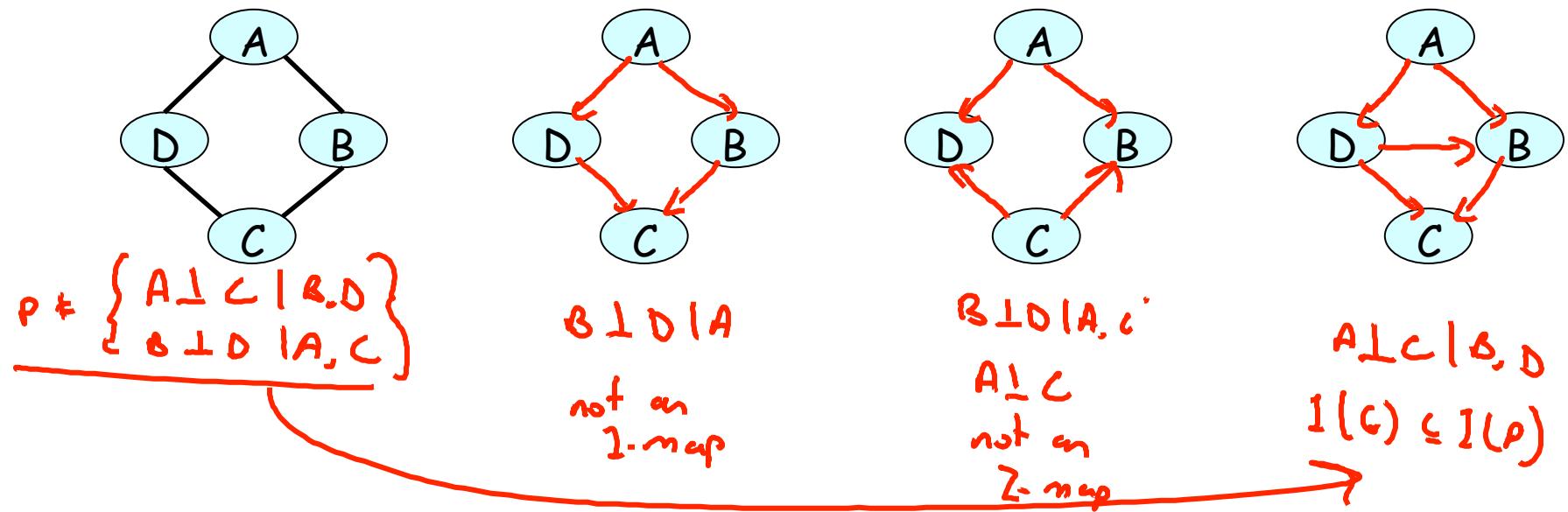
- Minimal I-map: I-map without redundant edges
~~($\phi \rightarrow \psi$) $P(\psi | \phi) = P(\psi | \neg \phi)$~~
- Minimal I-map may still not capture $I(P)$



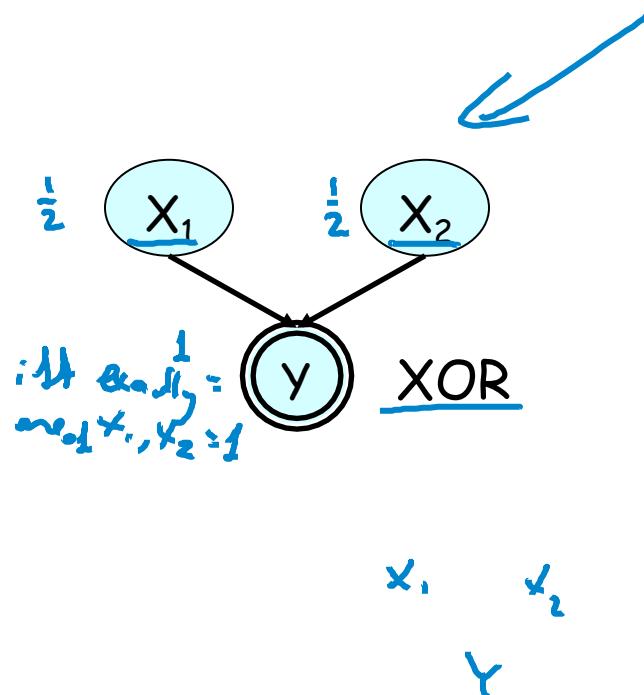
Perfect Map

- Perfect map: $\underline{I(G)} = \underline{I(P)}$
 - G perfectly captures independencies in P

Perfect Map



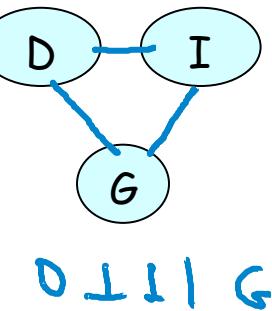
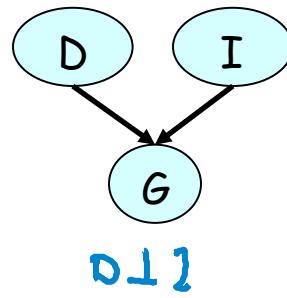
Another imperfect map



x_1	x_2	<u>y</u>	Prob
0	0	0	0.25
0	1	1	0.25
1	0	1	0.25
1	1	0	0.25

MN as a perfect map

- Perfect map: $I(\underline{H}) = I(P)$
 - H perfectly captures independencies in P

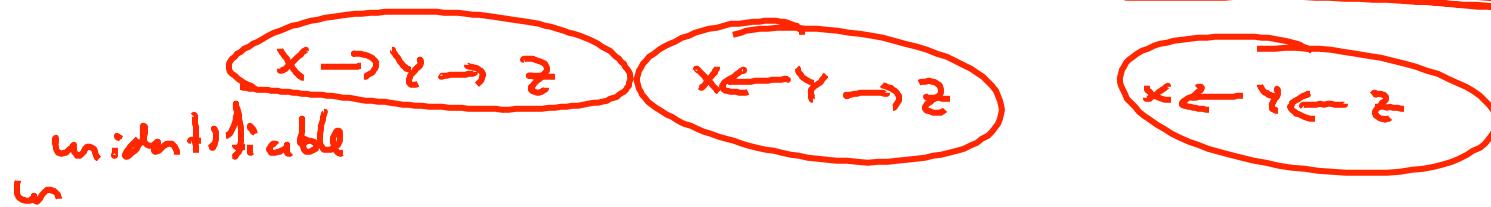


Uniqueness of Perfect Map

$G_1: X \rightarrow Y \quad I(G_1) = \emptyset$
 $G_2: Y \leftarrow X \quad I(G_2) = \emptyset \Rightarrow$ can represent
exactly the same
distribution

I-equivalence

Definition: Two graphs G_1 and G_2 over X_1, \dots, X_n are I-equivalent if $I(G_1) = I(G_2)$



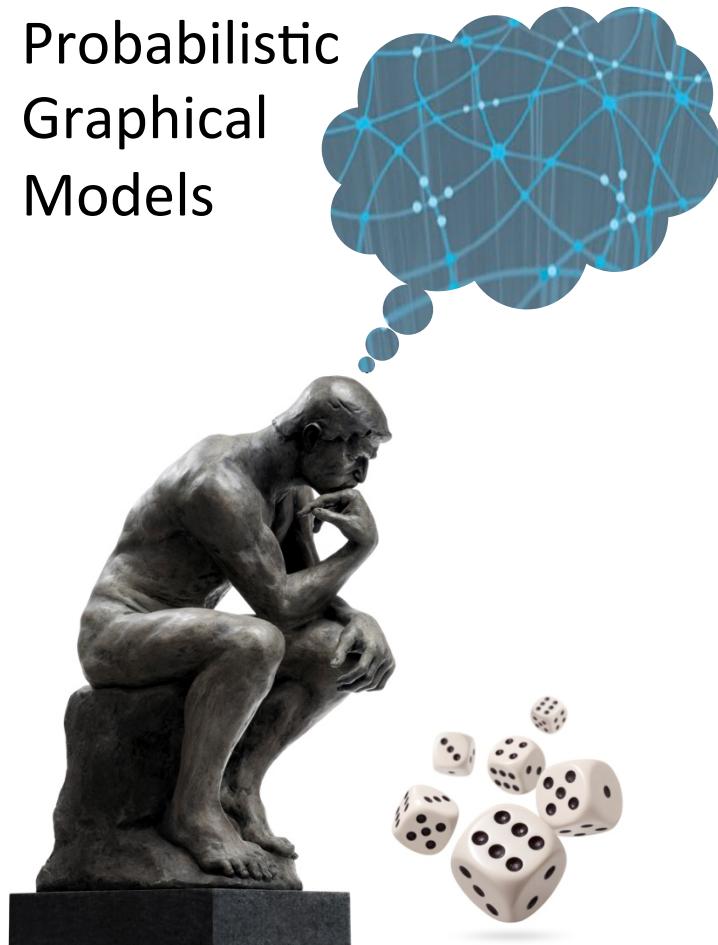
Most G 's have many I-equivalent variants

Summary

- Graphs that capture more of $I(P)$ are more compact and provide more insight
- A minimal I-map may fail to capture a lot of structure even if present *and representable as a PGM*
- A perfect map is great, but may not exist
- Converting BNs \leftrightarrow MNs loses independencies
 - BN to MN: loses independencies in v-structures
 - MN to BN: must add triangulating edges to loops



Probabilistic
Graphical
Models



Representation

Local Structure

Log-Linear
Models

Log-Linear Representation

$$\tilde{P} = \prod_i \phi_i(D_i)$$

$$\tilde{P} = \exp \left(- \sum_j \underbrace{w_j f_j(D_j)}_{\text{features}} \right)$$

$$\tilde{P} = \prod_j \underbrace{\exp (-w_j f_j(D_j))}_{\text{factor}}$$

- Each feature f_j has a scope $\underline{D_j}$
- Different features can have same scope

Representing Table Factors

$$\phi(X_1, X_2) = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$$

$$\begin{aligned} f_{12}^{00} &= \mathbf{1}\{X_1 = 0, X_2 = 0\} \\ f_{12}^{01} &= \mathbf{1}\{X_1 = 0, X_2 = 1\} \\ f_{12}^{10} &= \mathbf{1}\{X_1 = 1, X_2 = 0\} \\ f_{12}^{11} &= \mathbf{1}\{X_1 = 1, X_2 = 1\} \end{aligned}$$

General representation

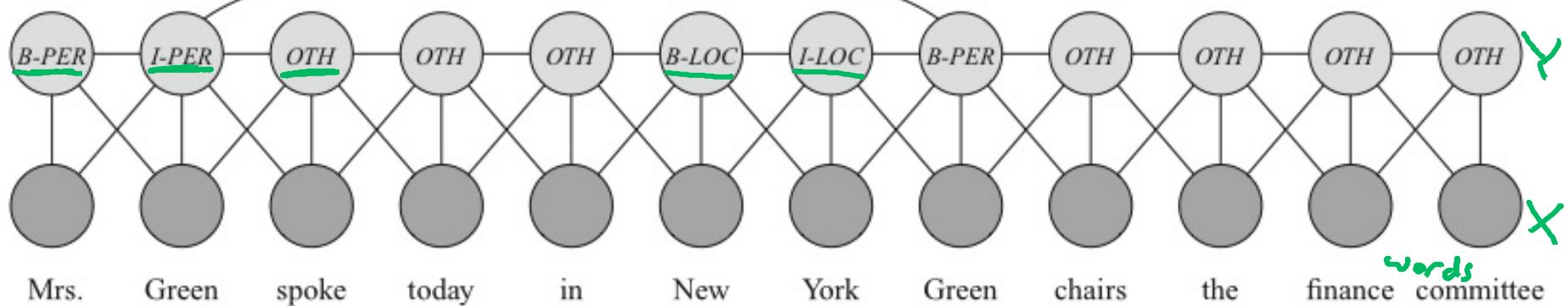
$$\phi(X_1, X_2) = \exp\left(-\sum_{kl} w_{kl} f_{ij}^{kl}(X_1, X_2)\right)$$

$$w_{kl} = -\log a_{kl}$$

$\exp(-w_{00})$ when $x_1=0, x_2=0$
 $\exp(-w_{0,1})$ when $x_1=0, x_2=1$

Daphne Koller

Features for Language



Features: word capitalized, word in atlas or name list, previous word is "Mrs", next word is "Times", ...

$$f(x_i, x_{i+1}) = \begin{cases} 1 & \{x_i = \text{person}, x_{i+1} \text{ is capitalized}\} \\ 1 & \{x_i = \text{B-loc}, x_{i+1} \text{ appears in Atlas}\} \end{cases}$$

Ising Model

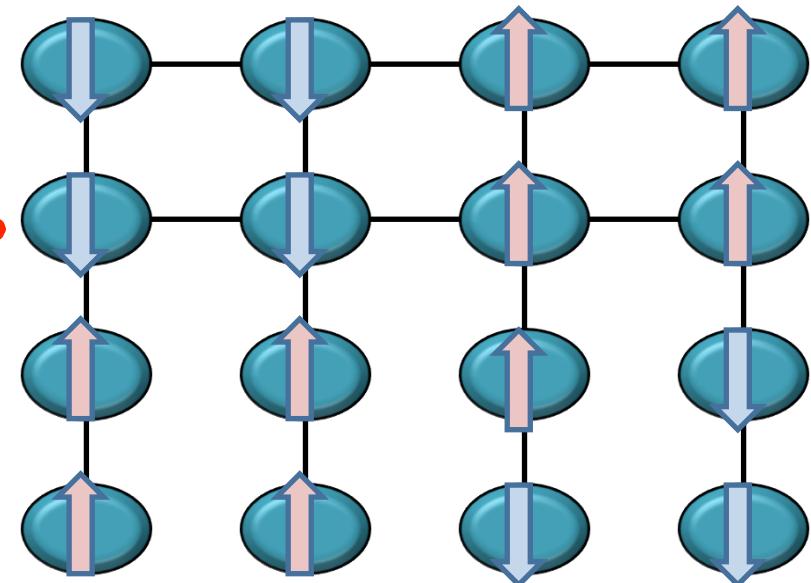
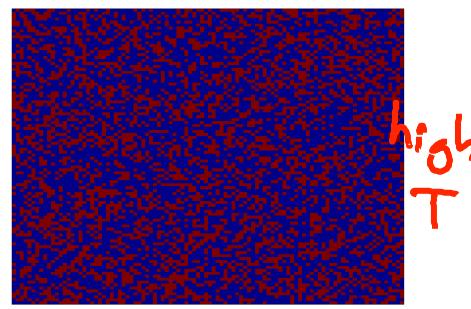
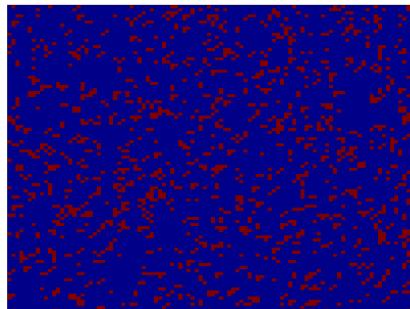
$$E(x_1, \dots, x_n) = - \sum_{i < j} w_{i,j} \cancel{x_i x_j}^{\text{conflict}} - \sum_i u_i x_i$$

pairwise joint spins

$x_i \in \{-1, +1\}$, $f_{i,j}(X_i, X_j) = \underline{\underline{X_i \cdot X_j}}$

law $P(X) \propto e^{-\frac{1}{T} E(X)}$

T groups $\frac{w_{ij}}{T} \rightarrow 0$
 T decreases



Metric MRFs

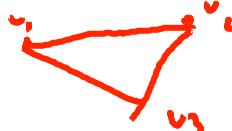
- All X_i take values in label space V



want X_i and X_j to take "similar" values

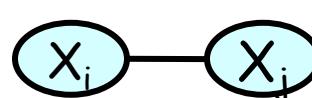
- Distance function $\mu : V \times V \rightarrow \mathbb{R}^+$

- Reflexivity: $\mu(v, v) = 0$ for all v
- Symmetry: $\mu(v_1, v_2) = \mu(v_2, v_1)$ for all v_1, v_2
- Triangle inequality: $\mu(v_1, v_2) \leq \mu(v_1, v_3) + \mu(v_3, v_2)$ for all v_1, v_2, v_3



Metric MRFs

- All X_i take values in label space V



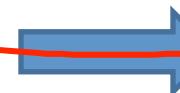
want X_i and X_j to
take "similar" values

- Distance function $\mu : V \times V \rightarrow \mathbb{R}$

$$\underline{f_{i,j}(X_i, X_j)} = \mu(X_i, X_j)$$

$$\exp(-w_{ij} f_{ij}(X_i, X_j))$$

w_{ij}
values of X_i and X_j far in μ



$$w_{ij} > 0$$

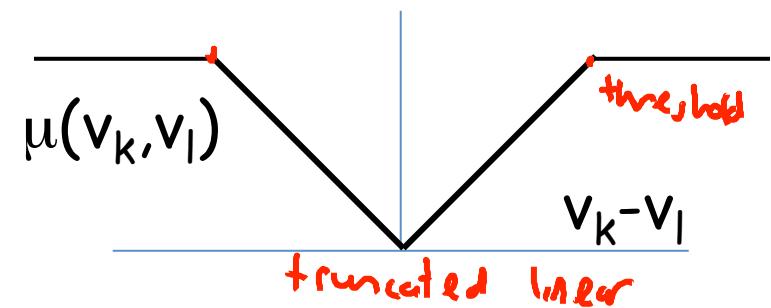
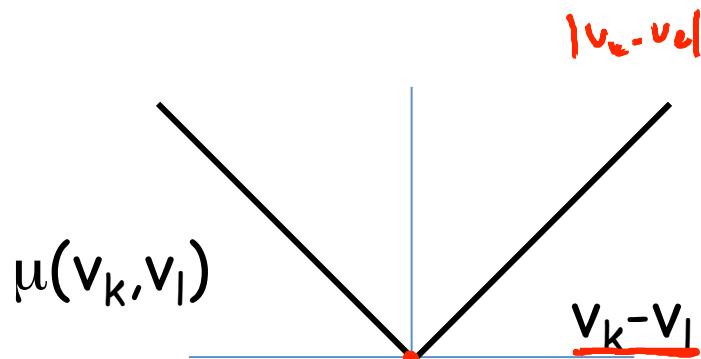
lower distance
higher probability

lower probability

Metric MRF Examples

$$\mu(v_k, v_l) = \begin{cases} 0 & v_k = v_l \\ 1 & \text{otherwise} \end{cases}$$

0	1	1	1
1	0	1	1
1	1	0	1
1	1	1	0

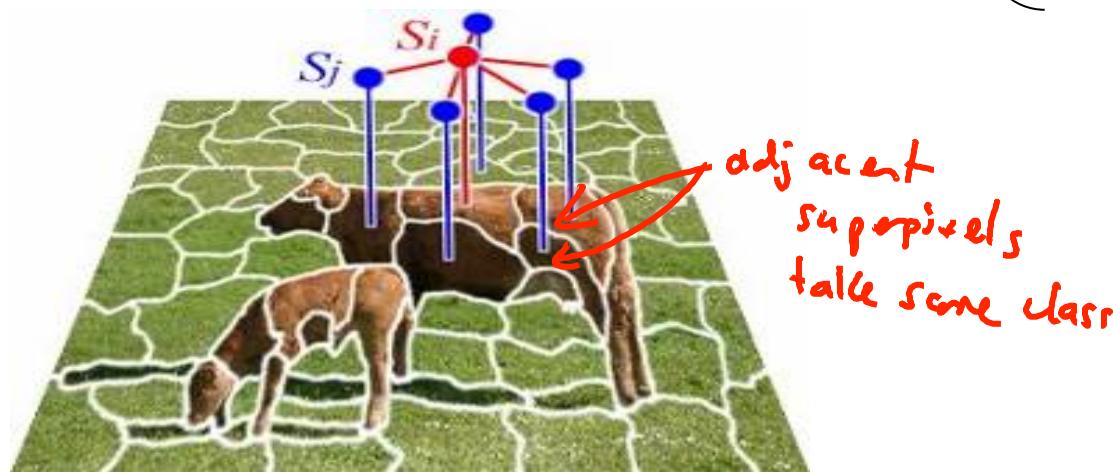


Daphne Koller

Metric MRF: Segmentation

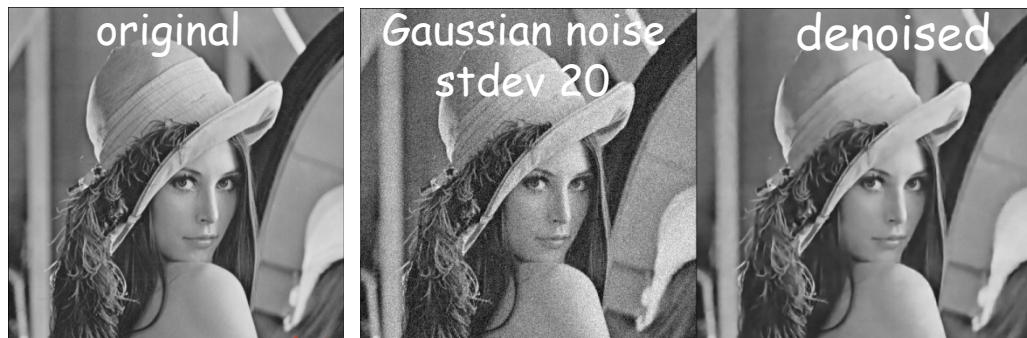
$$\mu(v_k, v_l) = \begin{cases} 0 & v_k = v_l \\ 1 & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$



Daphne Koller

Metric MRF: Denoising



$$\mu(v_k, v_l) = |v_k - v_l|$$

v_k - v_l

X noisy pixels

Y clean pixels

*Y: close to x,
Y: close to its neighbors*



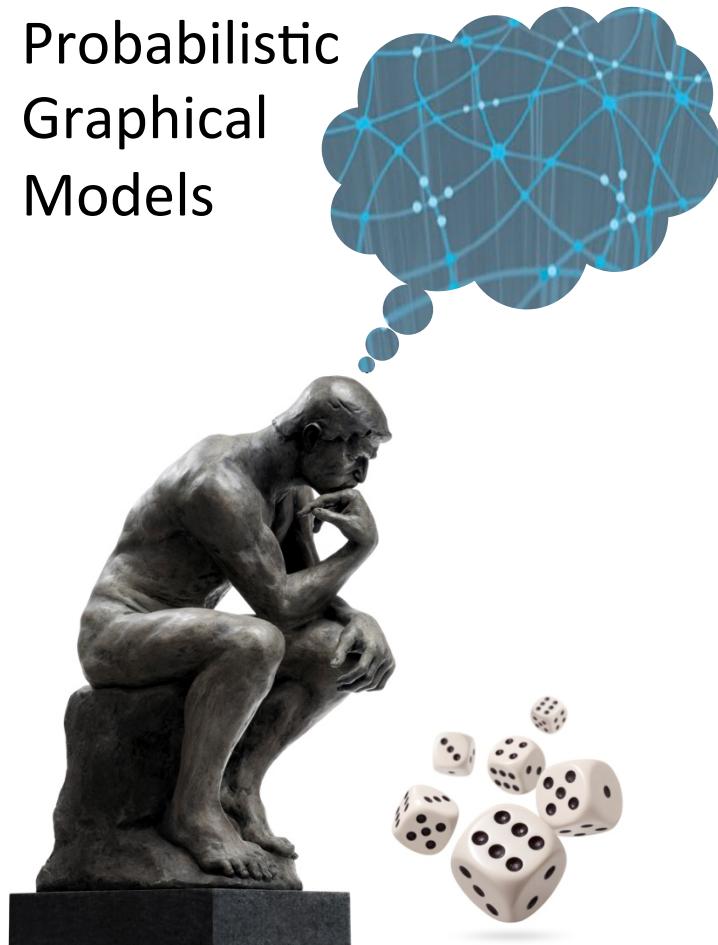
$$\mu(v_k, v_l) = \min(|v_k - v_l|, d)$$

v_k - v_l

Similar idea for stereo reconstruction

Daphne Koller

Probabilistic
Graphical
Models



Representation

Template Models

Shared
Features in Log-
Linear Models

Ising Models

- In most MRFs, same feature and weight are used over many scopes

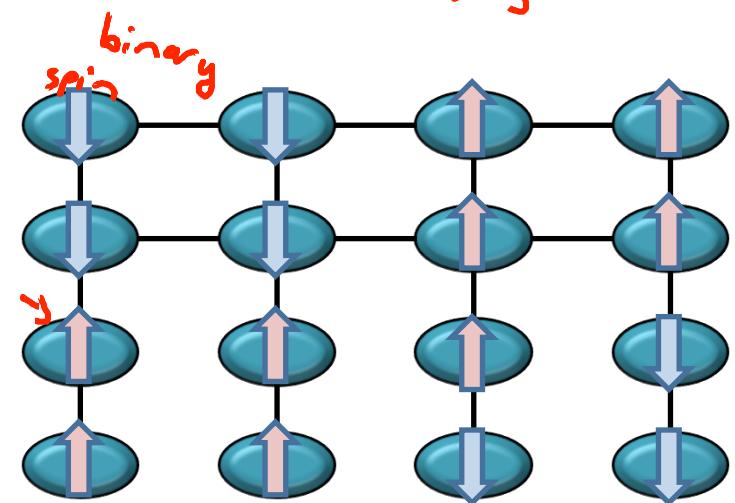
$x_i \in \{-1, +1\}$

Ising Model

$$E(x_1, \dots, x_n) = - \sum_{(i,j) \in \text{Edges}} w_{i,j} x_i x_j - \sum_i u_i x_i$$

w_{i,j} f(x_i, x_j)
weight same feature

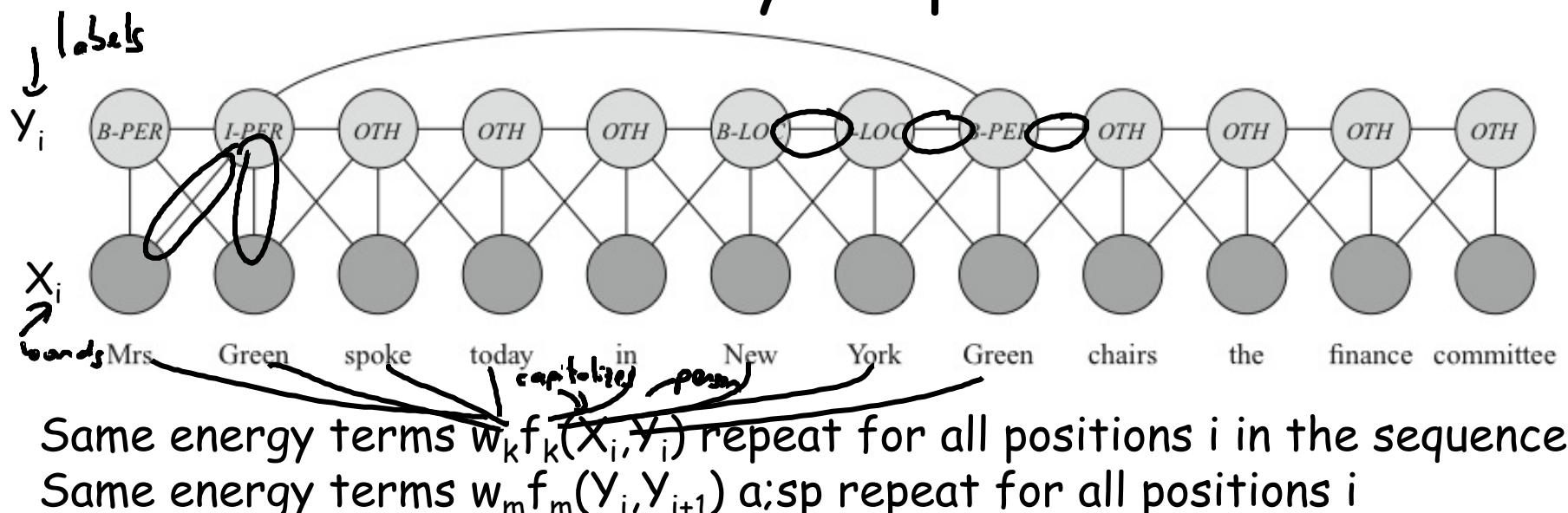
same weight for every adjacent pair



Daphne Koller

Natural Language Processing

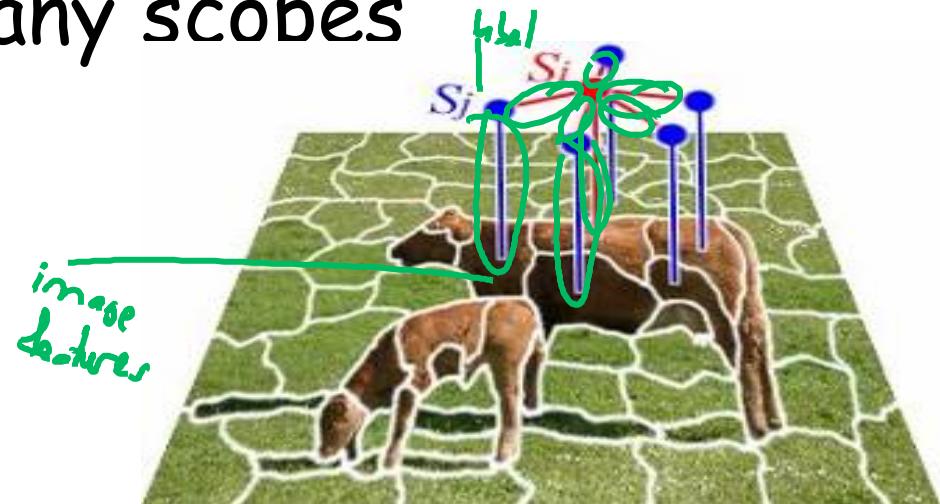
- In most MRFs, same feature and weight are used over many scopes



Daphne Koller

Image Segmentation

- In most MRFs, same feature and weight are used over many scopes



Same features and weights for all superpixels in the image

Repeated Features

- Need to specify for each feature f_k a set of scopes Scopes[f_k]
- For each $D_k \in \text{Scopes}[f_k]$ we have a term $w_k f_k(D_k)$ in the energy function

$$w_k \sum_{D_k \in \text{Scopes}(f_k)} f_k(D_k)$$

Summary

- Same feature & weight can be used for multiple subsets of variables
 - Pairs of adjacent pixels/atoms/words
 - Occurrences of same word in document
- Can provide a single template for multiple MNs
 - Different images
 - Different sentences
- Parameters and structure are reused within an MN and across different MNs
- Need to specify set of scopes for each feature